

(19) World Intellectual Property Organization
International Bureau

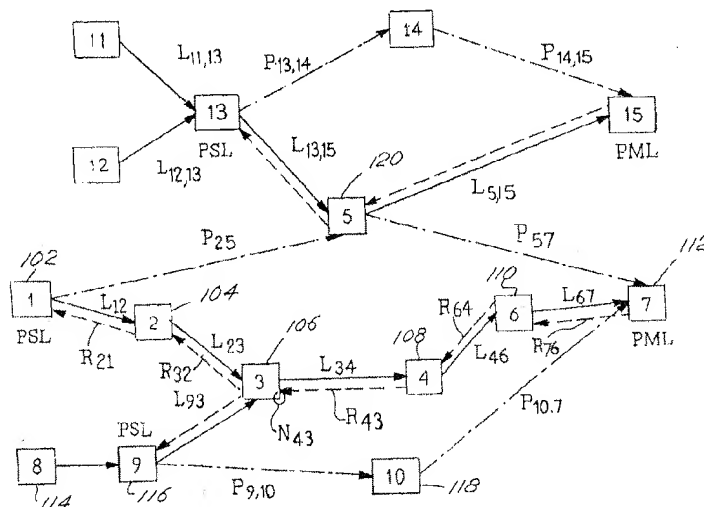


(43) International Publication Date
26 April 2001 (26.04.2001)

PCT

(10) International Publication Number
WO 01/29685 A1

- (51) International Patent Classification⁷: **G06F 15/173**
- (21) International Application Number: PCT/US00/28992
- (22) International Filing Date: 20 October 2000 (20.10.2000)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
60/160,840 21 October 1999 (21.10.1999) US
60/161,277 25 October 1999 (25.10.1999) US
60/187,798 8 March 2000 (08.03.2000) US
- (71) Applicant: **TELLABS OPERATIONS, INC.** [US/US];
4951 Indiana Avenue, Lisle, IL 60532 (US).
- (72) Inventors: **OWENS, Kenneth, R.**; 1106 Fourth Street, St. Louis, MO 63126 (US). **MAKAM, Srinivas, V.**; 1712 Ada Court, Naperville, IL 60540 (US). **HUANG, Changchen**; Apartment 303, 2833 Buckingham Drive, Lisle, IL 60532 (US). **VISHAL, Sharma**; 53 Pond Street, Belmont, MA 02478 (US).
- (54) Title: **METHOD AND APPARATUS FOR DETECTING MPLS NETWORK FAILURES**
- (81) Designated States (*national*): AE, AG, AL, AU, BA, BB, BG, BR, BZ, CA, CN, CR, CU, CZ, DM, DZ, EE, GD, GE, HR, HU, ID, IL, IN, IS, JP, KR, LC, LK, LR, LT, LV, MA, MG, MK, MN, MX, NO, NZ, PL, RO, SG, SI, SK, TR, TT, UA, UZ, VN, YU, ZA.
- (84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).
- Published:**
— With international search report.
— Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.



(57) Abstract: Recovery time upon the failure of a link (L12, L23, L34, ...) or switching system (102, 104, 106, 108, ...) in an asynchronous data network can be minimized if downstream data switches (102, 104, 106, 108, ...) can provide upstream messages (R21, R32, R43, ...) to upstream switching systems (102) indicating to upstream switching systems (102) that the downstream traffic arrived intact and was properly handled. In one embodiment, the loss or failure of the upstream status message to an upstream switching system (102), causes an upstream switching system to reroute data traffic around a failed switch (106) with a minimal amount of lost data.

WO 01/29685 A1

B2

**METHOD AND APPARATUS FOR DETECTING MPLS NETWORK
FAILURES**

RELATED APPLICATIONS

5 This application claims priority under 35 U.S.C. §119(e) to provisional applications serial nos. 60/160,840, filed October 21, 1999, 60/161,277 filed October 25, 1999 and 60/187,798 filed March 8, 2000, the entire writing and content of which is incorporated by reference.

FIELD OF THE INVENTION

10 This invention relates to data networks. In particular this invention relates to a method and apparatus for automatically routing a message upstream through a multi-protocol label switching (MPLS) network so as to trigger the re-routing of data onto an alternate path through the MPLS network.

BACKGROUND OF THE INVENTION

15 Multiprotocol Label Switching (MPLS) is a new technology that combines OSI layer 2 switching technologies and OSI layer 3 routing technologies. The advantages of MPLS over other technologies include the flexible networking fabric that provides increased performance and scalability. This includes Internet traffic engineering aspects that include Quality of Service (QoS)/Class of Service (COS) and facilitate the use of Virtual Private
20 Networks (VPNs).

 The Internet Engineering Task Force (IETF) defines MPLS as a standards-based approach to applying label switching technology to large-scale networks. The IETF is defining MPLS in response to numerous interrelated problems that need immediate attention. These problems include, scaling IP networks to meet the growing demands of Internet traffic,
25 enabling differentiated levels of IP-based services to be provisioned, merging disparate traffic types onto a single IP network, and improving operational efficiency in a competitive environment.

 The key concept in MPLS is identifying and marking IP packets with labels and forwarding them to a modified switch or router, which then uses the labels to switch the
30 packets through the network. The labels are created and assigned to IP packets based upon the information gathered from existing IP routing protocols.

 The label stack is represented as a sequence of "label stack entries". Each label stack entry is represented by 4 octets.

```

0      1      2      3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+ Label
|    Label                | Exp |S|   TTL   | Stack
+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+--+ Entry
```

Label: Label Value, 20 bits
Exp: Experimental Use, 3 bits
S: Bottom of Stack, 1 bit
TTL: Time to Live, 8 bits

The label stack entries appear after the data link layer headers, but before any network layer headers. The top of the label stack appears earliest in the packet, and the bottom appears latest. The network layer packet immediately follows the label stack entry which has the S bit set.

Multi-protocol label switching (MPLS) networks are typically comprised of several packet-based switching systems interconnected by a variety of media (e.g., coaxial or fiber optic cable, unshielded twisted pair or even point-to-point microwave wireless) in a mesh-topology network similar to the public switched telephone network. In such a network, there might be several paths through the network between any two endpoints. MPLS networks carry data as packets wherein each packet includes a label on identifying a switched path through the network. The data label is appended to data packets so as to define a pathway through the network over which the data packets are to be routed.

A problem with any data network, including an MPLS network, is the amount of time required to recover from either a link failure or a switch failure. Empirical data shows that the time required to recover from a network failure can take several seconds to several minutes, an unacceptably long time. A method and apparatus by which the recovery time for a link or switch failure can be reduced to perhaps less than a few hundred milliseconds would be a significant improvement over the prior art fault recovery mechanisms used on MPLS networks to date. A method and apparatus by which a switch over from a working path to a protection path would facilitate MPLS network reliability.

SUMMARY OF THE INVENTION

In an MPLS data network comprised of various transmission media linking various types of switching systems, network fault recovery time is reduced by using a reverse-

directed status message that is generated by a data switch that is down-stream from a switching system from which data is received. The reverse-directed or upstream status message is sent over a pre-determined pathway (i.e. through pre-determined switches and/or over pre-determined data links) which originates from a destination switch or node in an MPLS network to upstream switching systems. This so-called reverse notification tree carries a message or messages that are used to indicate the functionality (or non-functionality) of the downstream switch, switches or links of the MPLS network. As long as an upstream MPLS switching system continues to receive the reverse-directed status message from a downstream switch via the reverse notification tree, the switching systems that receive such a message consider the downstream switch and pathways to be intact. Accordingly, data packets continue to be sent downstream for subsequent routing and/or processing. If the reverse-directed status message is lost or discontinued, either because of a switch failure or a link failure, the upstream switching system considers the downstream switch or link to have failed and thereafter begins executing a procedure by which data is rerouted over an alternate data path through the network. In the preferred embodiment, the alternate data path over which downstream information is sent is a pre-established protection path and is known to a protection switch in advance, thereby minimizing data loss attributable to the time it might take to calculate a dynamic alternate protection path.

Switches in the network and their interconnections can be modeled using a directed acyclical graph by which a downstream switch knows the identity of the upstream switch to which the failure notice should be sent. In the preferred embodiment, at least one upstream switch routing the MPLS data re-directs data onto a protection path through the network between the same two endpoints by using messages carried over a different pathway, discussed hereinafter and referred to as a reverse notification tree. By way of the reverse notification tree, data loss caused by either a link or switch failure can be minimized by the prompt rerouting of the data through an alternate or recovery data path through the network.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a simplified block diagram of an MPLS protection configuration.

Figure 2 depicts exemplary message flows in an MPLS network.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Figure 1 shows a simplified block diagram of a packetized-data switching network 100. Each of the squares shown in Figure 1 including boxes represented by reference numerals 102, 104, 106, 108, 110, 112, 114, 116, 118 and 120 represent one or more types of asynchronous switching systems that asynchronously receive data in e.g., packets, cells or

frames from an “upstream” switch and route, direct, couple or otherwise send the data onward to another “downstream” switch logically closer to the ultimate destination for the data. By way of example, these switching systems might be internet protocol (IP) routers, asynchronous transfer mode (ATM) switches, frame relays switches or other types of packetized-data switching systems implemented to receive packetized data over a transmission line and reroute the data onto one or more output ports to which are connected transmission media coupled to other switching systems.

In Figure 1, switching system number 1, (identified by reference numeral 102) is coupled to another switching system, no. 2, (represented by reference numeral 104) and switching system no. 5 (represented by reference numeral 120) via links L_{12} and P_{25} respectively. Switching system no. 2 and no. 5 are “downstream” from no. 1; no. 1 is considered to be “upstream” from switch no. 2 and no. 5.

Similarly switching system no. 3, (represented by reference numeral 106) is coupled to switching systems 2, 4 and 9 (represented by reference numerals 104, 108 and 116 respectively) via transmission links L_{23} , L_{34} , and L_{93} respectively.

In routing data between switch no. 1 (represented by reference numeral 102) and switch no. 7 (represented by reference numeral 112) data might be routed between these two endpoints through a “*primary*” path that is comprised of links that logically or physically couple switches 2, 3, 4, 6 and 7 (identified by reference numerals 104, 106, 108, 110 and 112 respectively). The physical or logical links of the primary path between the endpoints which is 1 and 7 are represented by the vectors designated L_{12} , L_{23} , L_{34} , L_{46} and L_{67} . This path is known in the art as the working or primary path through the network. The links of the various paths shown in Figure 1 (represented by the vectors L_{12} , L_{23} , L_{34} , L_{46} and L_{67}), and therefore the paths themselves, might be constructed of direct pathways (e.g., fiber optic cable, coaxial cable, unshielded twisted pairs of copper wires, or microwave radio) between the various switches. Alternate embodiments of the paths or links between switches of the network of Figure 1 would also include using direct pathways, and intermediate switches or switch networks, (not shown in Figure 1, but still part of the path or link coupling one or more switching systems to another). By way of example and not of limitation, the data switches shown in Figure 1 might be IP switches but such IP switches could be linked together using one or more ATM switches or ATM networks.

The MPLS Protection Path

In an MPLS network, there is almost always a “*protection*” path, which is an alternate path through the network linking two endpoints. The protection path entry and exit

points are usually accessible to only *protection* switches. A protection switch is a switch that can re-route traffic onto a protection pathway. Like the other links described above, a protection pathway can be comprised of direct data paths, but also switches or switching systems, by which data can be sent through a network between two or more endpoints.

5 In an MPLS network, a protection path is set up using at least one protection switch element so as to be able to carry data from a source to a destination in the event the primary path or switch thereof fails for one reason or another. The operation of a protection switch is shown in Figure 1 by way of example.

10 In Figure 1, a working path between switch 1 and 7 exists through switches 1, 2, 3, 4, 6 and 7 and the links between the switches. A protection path for the portion of the working path that runs through switches 2, 3, 4 and 6 is the path designated by links P₂₅ and P₂₇ and which runs through switch 5, (identified by reference numeral 120). The protection path extends between endpoint switches 1 and 7 but through only switch 5 (identified by reference numeral 120). Alternate embodiments of a protection path might extend through multiple
15 switches. In the network 100, either a link or switch loss between switch 1 and 7 can be overcome by re-routing traffic for switch 2 through switch 5 instead. Switch 5 then routes the data to switch 7. Switch 1 (identified by reference numeral 102) is considered to be a protection switch element.

20 Another working path between switch 8 and switch 7 of the network 100 (identified by reference numerals 114 and 112 respectively) exists through switches 9, 3, 4, 6 & 7 (identified by reference numerals 116, 106, 108 and 110 respectively) and the links between them. A protection path for data from switch 8 (reference numeral 114) to switch 7 (reference numeral 112) through the network 100 exists via switch 10, (reference numeral 118) such that if data traffic from switch 8 (reference numeral 114) is lost somewhere between switch 9
25 (reference numeral 116) and switch 7 (reference numeral 112), switch 9 can re-route such data traffic to switch 10 (reference numeral 118). Switch 10 can then route such data to switch 7. Switch 9 is therefore considered to be a protection switch element.

30 If an alternate data path, i.e. a protection path, is pre-determined, i.e. set up or established in advance, data loss attributable to a switch or link failure can be minimized. If a protection switch is pre-programmed to re-route data upon its receipt of an appropriate command or signal, the protection switch element can almost immediately start sending data to the proper destination via the protection path.

The Liveness Message

In the event of a pathway failure causing downstream data to be lost at a downstream switch, such as by either a switch failure or a link failure, anywhere along a primary or working path, a protection switch element (PSL), such as switch no. 1 (identified by reference numeral 102) can re-route data traffic through the protection path so as to have the data for the endpoint switch no. 7 delivered as quickly as possible to the endpoint at switch no. 7 (identified by reference numeral 112). The ability to re-route data to a protection path is made considerably more valuable if the decision to switch over to a protection path is based upon an affirmative notice that a switch over is needed. In one embodiment, this affirmative notice is in the form of an upstream liveness message, the loss of which indicates a pathway failure. As long as a liveness message is received at an upstream switch from a downstream switch, the upstream switch can assume that the pathway between the two switches is intact and that the downstream switch is functional.

The liveness message is sent upstream, and is repeated, so long as downstream data is received. In one embodiment, the liveness message is repeated periodically, at some appropriate repetition rate, however, an alternate embodiment includes sending an upstream message aperiodically, and also without regard to the time between messages. Downstream data loss at a switch will trigger the switch to inhibit upstream-directed liveness messages.

In another embodiment, a predetermined downstream liveness message is sent instead of, or in addition to downstream data. The receipt of a such a downstream liveness message will cause a recipient switch to maintain its generation of upstream liveness messages. For purposes of claim construction, downstream liveness messages and downstream voice or downstream data (considered to be interchangeable for claim construction purposes) are considered to be equivalent under the rubric of "a first data message" or a "first message," the reception of which can either inhibit an upstream liveness message, or in alternate embodiments, trigger an upstream liveness message.

Those skilled in the art will appreciate that while the loss of a liveness message can trigger a protection switch of data to a protection path, an equivalent alternate embodiment of the invention includes a method by which a protection switchover is initiated or occurs upon the receipt of either a liveness message or a notification message to a protection switch element. In such a case, the liveness message and/or the notification acts as a trigger event to cause information to be re-routed onto the protection path.

In the event of a switch or link failure anywhere between the path endpoint switches 1 and 7, data re-routing is accomplished faster by using a reverse-directed status message that

is sent backward or upstream toward the protection switch element no. 1 (reference numeral 102) by one or more of the switches 2, 3, 4, 6 or 7 (reference numerals 104, 106, 108, 110 or 112) of the primary pathway, links L_{12} , L_{23} , L_{34} , L_{46} , L_{67} . In the preferred embodiment this reverse direction data message is known as a "liveness message" the format of which is a design choice and dependent upon nature of the switches of the network 100, but the function of which is to indicate to upstream switches that the data traffic sent to the downstream switch arrived intact and on time.

The structure of a liveness message will vary depending upon whether the network switches are ATM, IP, Ethernet or other types of switches, as those skilled in the art will recognize. Unlike known fault detection methods, the liveness message is not a copy, or loop-back of the downstream data. The salient aspect of the liveness message is that it is an informational status message, preferably sent at periodic intervals between adjacent nodes, indicating the operational condition of the switch from which it was sent. As set forth above, alternate embodiments include sending a liveness message aperiodically. In at least one of the foregoing embodiments, the fact that the liveness message is received at an upstream switch (with respect to the downstream data) is evidence that the link between the switches, over which downstream data would be sent, is intact and that the switch that generated the liveness message is at least somewhat functional.

While the preferred embodiment contemplates that the liveness message is sent upstream from a switch, directly to the switch that sent the downstream data, alternate embodiments of the invention contemplate that a liveness message, or equivalent thereof, could be sent in both upstream and downstream directions, between other nodes, or from one switch to an intermediate transfer point, which for purposes of claim construction are considered to be equivalent embodiments. By way of example, with reference to Figure 1, switch no. 4 (identified by reference numeral 108) will send a liveness message, upstream to switch no. 3 (reference numeral 106) in response to data sent downstream (or perhaps a downstream liveness message, sent downstream) from switch no. 3 to switch no. 4. If the liveness message from switch no. 4 is lost by or not received by switch no. 3, switch no. 3 can infer that either the link R_{43} between the two switches failed, or switch no. 4 failed. If switch no. 3 was carrying data from switch no. 9 (reference no. 116) and which is a protection switch element having access to a protection path, switch no. 3 would need to inhibit its liveness message to switch no. 9 or generate an error message to switch no. 9, thereby instructing switch no. 9 to re-route traffic from switch no. 3, to the protection path through switch no. 10 (reference numeral 118)

As for data routed through switch no. 3 that comes from switch no. 2 (reference numeral 104), a liveness message loss from switch no. 4 will require switch no. 3 to inhibit the liveness message to switch no. 2, or send an error message to switch no. 2. This procedure is then repeated to switch no. 1, instructing switch no. 1 to make a protection switch through switch no. 5 (reference no. 120).

When a liveness message is lost, its failure is considered to be indicative of a path failure of either a link or a switch. Still other embodiments of the invention contemplate sending a downstream liveness message, sent from an upstream switch to a downstream switch thereby indicating to a downstream switch that the upstream switch and link are functional.

As set forth above, the format of a liveness message will depend upon the type of switching systems used in the network. IP switches and ATM switches will need to comply with their respective protocols. Alternative embodiments of the invention would certainly contemplate other sorts of liveness messages having different formats with the salient feature of the message being that the message indicates to an upstream switch that downstream directed data messages were received by a downstream switch intact.

In Figure 1, the links over which reverse notification status messages (i.e. the upstream liveness messages) are sent, are designated by the reverse directed vectors, one of which is shown in Figure 1 (and identified by reference numeral R_{76}). By way of example if link L_{67} should fail causing a data loss to the endpoint switch no. 7, the corresponding loss of the liveness message ordinarily sent from switch 7 to switch 6 would provide an indication to switch no. 6 that either the link or the switch 7 failed whereupon switch no. 6 would begin notifying the protection switch (switch no. 1) upstream from it by way of a reverse notification message sent on reverse link R_{64} that would be sent to switch no. 4, (represented by reference numeral 108). Similarly, switch no. 4 would thereafter return a reverse notification message on reverse link R_{43} to switch no. 3. Switch no. 3 returns another reverse notification message on reverse link R_{32} to switch 2 which then returns a reverse notification message on reverse link R_{21} to the origination node 1.

The ultimate destination of the upstream message, and in this case the reverse notification message, is a switching node (i.e. a switch or switching system) that is capable of re-routing downstream traffic, data or messages onto a different path, i.e., a protection path, usually comprised of at least a different transmission route, possibly including a different transmission media as well (coax to fiber; fiber to microwave etc.). Whether the upstream message goes through another switch on its way to the switching node (which has the

capability of re-routing data to the protection path) or is directly sent to the switching node from a downstream switch around an intermediate switch (for example, sending a liveness message directly from switch 6 to switch 1) would still provide equal functionality in that the switching node will eventually receive notification that it needs to re-route traffic, data or message onto the protection path. Sending the aliveness message directly to the protection switch or routing the aliveness message via intervening switches are considered to be equivalent embodiments for purposes of claim construction.

Inasmuch as switch no. 1 in Figure 1 is designated as a "protection switch element" meaning that it is coupled to and capable of routing data onto a protection path P_{25} , the protection switch element 1 (identified by reference numeral 102) reroutes traffic to switch no. 7 via a protection path designated by P_{25} and P_{57} and that runs through switch no. 5 (identified by reference numeral 120).

In the preferred embodiment, the switches of the network maintain tables of network switches upon which incoming data is received and a table of network switches to which outgoing data is routed. By keeping a record of where outgoing data from a switch originates from, it is possible for a switch of the network 100 to promptly notify an upstream switch of a downstream link or switch failure.

In the process described above, each of the switches of the network sequentially notifies at least one switch upstream from it. Alternate (and for purposes of claim construction, equivalent) embodiments of the invention could certainly provide upstream notification messages directly from any downstream switch to every other upstream switch in a pathway. In such an embodiment, switch no. 6 might send a reverse notification message directly to the protection switch element 1 via a direct link thereby notifying the protection switch to immediately reroute data to the protection path P_{27} and P_{57} via switch no. 5. Switch no. 6 might also send a reverse notification (liveness) message to the other switching systems of the network as well.

The Reverse Notification Tree

The implementation of the upstream notification message, and its conveyance upstream to a protection switch element, is enabled by using an upstream pathway denominated herein as a reverse notification tree or "RNT." The RNT is an "upstream" signal pathway that allows messages from a protection path end point to be sent "upstream" to one or more protection path (and working path) starting point switches, nodes or starting points. In the preferred embodiment, the RNT passes through the same switches and over or through the links that comprise the working path (albeit over different transmission media) and for

claim construction purposes can be considered to be “coincident” with the working path. Alternate embodiments of the invention would include a reverse notification tree that runs through one or more switches or nodes that are not part of the working path. or which are only partly “coincident.”

5 With respect to Figure 1, node 7, identified by reference numeral 112, is the RNT starting point or head end. Nodes 1 and 9, which are identified by reference numerals 102 and 116, are the end points of the RNT and to which upstream protection switch messages would be sent from any node or switch between nodes 1, 9 and 7. Intervening nodes 3, 4 and 6, identified by reference numerals 106, 108 and 110 respectively, are constituent elements or
10 parts of the RNT.

The RNT can be established in association with the working path(s) simply by making each switching system along a working path “remember” its upstream neighbor (or the collection of upstream neighbors whose working paths converge at a network switching element and exit as one). A table or other data structure stored in memory (such as RAM,
15 ROM, EEPROM, or disk) of the switches of the paths can be configured to store data identifying switches coupled to a switching system in, or part of a working path as well as a protection path.

With respect to the network shown in Figure 1, Table 1 below shows that incoming or “Ingress” RNT messages to switch no. 3 from switch 4 are labeled “N43” (not shown in
20 Figure 1) and that these messages arrive at switch no. 3 from switch 4 at an inbound or “Ingress” interface I34 (not shown in Figure 1). Because switch no. 3 receives downstream messages from two (2) different switches, (i.e. switch 2 and switch 9) both of these two upstream switches must be sent an upstream notification therefore requiring two separate upstream messages from switch 3. Upstream RNT messages to switch 2 are labeled “N32”
25 and appear or are sent from interface I23. Upstream RNT messages to switch 9 are labeled “N93” and are sent from interface I93.

A	B	C	D	E	F
Ingress Label of RNT	Ingress Interface of RNT	Egress Label of RNT	Egress Interface of RNT	Egress Label of RNT	Egress Interface of RNT
N43	I34	N32	I23	N39	I93

Table 1. An inverse cross-connect reverse notification tree table for Switch 3 of Figure 1.

5 The reverse path (upstream) to switch 3 from switch 4 is labeled N43; the switch 3 interface for this data is designated I34. An upstream message received at I34 and that is labeled N43, is sent out from switch 3, via the interfaces I23 and I93 and labeled N32 and N39 respectively.

10 Table 2 shows the egress and interface labels of the working or downstream path from switch 3 and the originating switches for that data.

The working path (downstream) path from switch 3 is to switch 4 and is labeled "L34." The switch 3 interface for this data is designated "I34." The data sent downstream from switch 3 originates from switch 2 and switch 9, which are referred to in Table 2 as "Next Hop" switches.

15 Switch no. 2 originates data to switch no. 3 and that data is received at switch no. 3 on interface "I2." Data from switch no. 9 is received at switch no. 3 at interface "I9." The RNT or upstream notification to switch no. 2 leaves switch no. 3 on its RNT interface "I23." RNT notification to switch no. 9 leaves switch no. 3 from "I93."

20	A	B	C	D	E	F
	Egress Label of Working Path	Egress Interface of Working Path	Next Hop IP Address of RNT	Egress Interface of RNT	Next Hop IP Address of RNT	Egress Interface of RNT
	L34	I34	I2	I23	I9	I93

Table 2. An inverse cross-connect table for a hop-by-hop reverse notification tree.

25

A fault on the link between switch 3 and 4 in the downstream direction can be detected at a downstream node, switch 4 perhaps, via either a downstream data loss, a downstream liveness message loss or via a path failure (PF) or link failure (LF). A path failure (PF) is a software-

detected hardware fault. A link failure (LF) is a hardware-detected hardware fault. Upon the detection of a fault, the downstream node (switch 4) will periodically transmit fault indication signal (FIS) messages to its upstream neighbor (switch 3) (via the uplink R_{43}), which will propagate further upstream (using its inverse cross-connect table) until the FIS messages eventually reach the appropriate Protection Switch Element (switch 1 and/or 9). Receipt of an FIS message will trigger the protection switch over of data from the working path to the protection path. From Table 1, messages received at switch no. 3 are labeled "N43." Therefore, in Fig. 1, if link L34 has a fault, switch 4 can detect the fault via either a lost downstream liveness message from switch no. 3, or by detecting a Link Failure (LF) or Path Failure (PF) and thereafter start transmitting an FIS packet back to switch 3 on link L_{23} as represented by a message on link R_{43} . From Tables 1 and 2, there are two egress messages and interfaces from switch no. 3, which identify the upstream switches that are to be "notified" of a failure downstream from switch no. 3. (The traffic in the queues of switch 3 will continue to be serviced.) By using similar tables, switch 2 in turn will propagate the FIS over the RNT back to switch 1. The actual protection switch will be performed by switch 1, after the receipt of the first FIS. Switch 3 will stop transmitting FIS messages "t" time units after the transmission of the first FIS message.

In the preferred embodiment, only one RNT is required for all the working paths that merge (either physically or virtually) to form the multipoint-to-point "forward" or "downstream" path. Figure 1 shows that at least two (2) working paths (one path of which is comprised of switch elements 1, 2 and 3 that are identified by reference numerals 102, 104 and 106; a second path of which is comprised of switch elements 8, 9 and 3 that are identified by reference numerals 114, 116 and 106) converge at switch element 3 (identified by reference numeral 106). Alternate (and for purposes of claim construction, equivalent) embodiments would include using multiple RNTs for a single working path that has multiple paths that converge at a single node (switches of each path that converges might form different RNTs) as well as using multiple RNTs for a single working path.

The RNT is rooted at an appropriately chosen label switched router ("LSR"), (which hereafter is referred to as an MPLS network switch element) along the common segment of the merged working paths and is terminated at the protection switch elements (PSLs). Intermediate network switching elements on the converged working paths typically share the same RNT reducing signaling overhead associated with recovery. Unlike schemes that treat each label switched path (LSP) independently, and require signaling between a protection

switch element and a destination switch individually for each LSP, the RNT allows for only one (or a small number of) signaling messages on the shared segments of the LSPs.

The RNT can be implemented either at Layer 3 or at Layer 2 of the OSI, 7-layer protocol stack. In either case, delay along the RNT needs to be carefully controlled. This may be accomplished by giving the highest priority to the fault and repair notification packets, which travel along the RNT. We can therefore have a situation where different protection domains share a common RNT.

A protection “domain” is considered to be the switches and links of both a working path and protection path. For example, in Fig. 1, the protection domain bounded by network switch element 1 and network switch element 7, is denoted by {1-2-3-4-6-7, 1-5-7}.

When different protection domains have different RNTs, two cases may arise, depending on whether or not any portions of the two domains overlap, that is, have nodes or links in common. If the protection domains do not overlap, the protection domains are considered to be independent. By virtue of the RNTs in the two domains being different, neither of the working paths nor the RNTs of the two domains can overlap. In other words, failures in one domain do not interact with failures in the other domain. For example, the protection domain defined by {9-3-4-6-7, 9-10-7} is completely independent of the domain defined by {11-13-5-15, 11-13-14-15}. As a result, as long as faults occur in independent domains, the network shown in Fig. 1 can tolerate multiple faults (for example, simultaneous failures on the working path in each domain). If protection domains with disjoint RNTs overlap, it implies that the protection path of one intersects the working path of the other. Therefore, although failures on the working paths of the two domains do not affect one another, failures on the protection path of one may affect the working path of the other and visa versa. For example, the protection domain defined by {1-2-3-4-6-7, 1-5-7} is not independent of the domain defined by {11-13-5-15, 11-13-14-15} since LSR 5 lies on the protection path in the former domain and on the working path in the latter domain. When protection domains have the same RNT, different failures along the working paths may affect both paths differently. As shown in Fig. 1, for example, working paths 1-2-3-4-5-7 and 9-3-4-6-7 share the same RNT. As a result, for a failure on some segments of the working path, both domains will be affected, resulting in a protection switch in both (for example, the segment 3-4-6-7 in Fig. 1). Likewise, for failures on other segments of the working path, only one domain may be affected (for example, failure on segment 2-3 affects only the first working path 1-2-3-4-6-7, where as failure on the segment 9-3 affects only the second working path 9-3-4-6-7).

There are a number of ways to establish a protection domain, i.e., a working path and a protection path through an MPLS network. Establishing a protection path first requires the identification of the working path (embodied as some series of switches and path links through the MPLS network from a sending node to a destination node). In most cases, the working path and its corresponding recovery path are specified during a network switch path or connection setup procedure, either via a path selection algorithm (running at a centralized location or at an ingress network switch element) or via an administrative configuration (e.g. a manual specification of switches that comprise the protection path).

The specification of either a protection or working path, does not, strictly speaking, require the entire path to be explicitly specified. Rather, it requires only that the head end node or switching node and end or destination switch or node (of the respective paths) be specified. In the absence of a destination switch/node specification, the path egress points out of the MPLS network or domain need to be specified, with the segments between them being – “loosely” determined or routed. In other words, a working path would be established between the two nodes at the boundaries of a protection domain via (possibly loose) explicit (or source) routing using LDP/RSVP [label distribution protocol/reservation protocol] signaling (alternatively, via constraint-based routing, or using manual configuration), as set forth more fully below.

Figure 2 depicts message flows between four (4) different switches of an MPLS network that employs the path protection techniques disclosed herein. Vertical axes of Figure 2, identified by reference numerals 202, 204, 206 and 208, represent switching elements (shown in Figure 1) of an MPLS network from which and to which various types of messages are received and sent respectively. Switch 202 is upstream from switches 204, 206 and 208. Switch 204 is upstream from switch 206 as switch 206 is upstream from switch 208.

Protection Path Establishment

A Protection Domain Path is established by the identification of a protection switch or node and an end point switch or node in the MPLS network. The protection switch element (“PSL”) initiates the setup of the working LSP and elements and the recovery LSP and elements. It is also responsible for storing information about which network switch elements or portions thereof have protection enabled, and for maintaining a binding between outgoing labels specifying the working path and the protection/recovery path. The latter enables the switchover to the recovery path upon the receipt of a protection switch trigger.

A “label distribution protocol” is a set of procedures by which one LSR (i.e., a network switch element) informs another of the label bindings it has made. “Label binding”

is a process by which a message to be sent from a source to a destination is associated with various labels between the nodes that lie along the way, between the source and destination. By way of example, in Figure 1, a message to be sent from switch 1 to switch 7 is associated or bound to travel to switch 7 through switch 2 by, or using, the label L_{12} that is first
5 associated with the message at, or by, switch 1. Switch 2 in turn associates messages labeled L_{12} as bound for switch 3 and re-labels them as L_{23} . Re-labeling messages (e.g. re-labeling a message received at switch 2 on L_{12} , as the same message that is output from switch 2 but on L_{23} and which is received at switch 3, to be re-labeled by switch 3 and output again as L_{34}) is known as "label binding." Two or more LSRs, (network switch elements) which use a label
10 distribution protocol to exchange label binding information are known as "label distribution peers" with respect to the binding information they exchange.

The label distribution protocol also encompasses any negotiations in which two, label distribution peers, need to engage in order to learn of each other's MPLS capabilities. This label distribution protocol is referred to as path establishment signaling. MPLS defines two
15 methods for label distribution. These two methods are: Label Distribution Protocol (LDP/CR-LDP) and ReSerVation Protocol (RSVP).

Both LDP/CR-LDP and RSVP allow a path to be setup loosely (wherein each node determines it's next hop) or explicitly (wherein each node has been given it's next hop). These two protocols can be extended, as disclosed herein and by equivalents thereof, to
20 provide a novel mechanism by which protection path establishment can be signaled and created. Accordingly, a "Protection" field can be defined, and added as an extension to the existing label request messages in LDP/CR-LDP, and path message in RSVP protocols. The destination or end point node in the MPLS network participates in setting up a recovery path as a merging network switch element. The destination or end point node learns, during a
25 signaling or working/protection path configuration process, which working and protection paths are merged to the same outgoing network switch element.

Hosts and routers that support both RSVP and Multi-Protocol Label Switching can associate labels with RSVP flows. When MPLS and RSVP are combined, the definition of a flow can be made more flexible. Once a label switched path (LSP) is established, the traffic
30 through the path is defined by the label applied at the ingress node of the LSP (label switched path). The mapping of a label to traffic can be accomplished using a number of different criteria. The set of packets that are assigned the same label value by a specific node are said to belong to the same forwarding equivalence class (FEC) and effectively define the "RSVP

flow." When traffic is mapped onto a label-switched path in this way, we call the LSP an "LSP Tunnel". When labels are associated with traffic flows, it becomes possible for a router to identify the appropriate reservation state for a packet based on the packet's label value.

A Path message travels from a sender to receiver(s) along the same path(s) used by the data packets. The IP source address of a Path message must be an address of the sender it describes, while the destination address must be the DestAddress for the session. These addresses assure that the message will be correctly routed through a non-RSVP cloud.

The format of an exemplary RSVP message with the Protection Object extension is:

```

<Path Message> ::=  <Common Header> [ <INTEGRITY> ]
                    <SESSION> <RSVP_HOP>
                    [ <TIME_VALUES> ]
                    [ <EXPLICIT_ROUTE> ]
                    [ <PROTECTION> ]      /* The new message field. */
                    <LABEL_REQUEST>
                    [ <SESSION_ATTRIBUTE> ]
                    [ <POLICY_DATA> ... ]
                    <sender descriptor>

```

Label Distribution Protocol (LDP) is defined for distribution of labels inside one MPLS domain. One of the most important services that may be offered using MPLS in general, and LDP in particular, is support for constraint-based routing of traffic across the routed network. Constraint-based routing offers the opportunity to extend the information used to setup paths beyond what is available for the routing protocol. For instance, an LSP can be setup based on explicit route constraints, QoS constraints, and other constraints.

Constraint-based routing (CR) is a mechanism used to meet Traffic Engineering. These requirements may be met by extending LDP for support of constraint-based routed label switched paths (CR-LSPs).

The Path Vector TLV is used with the Hop Count TLV in Label Request and Label Mapping messages to implement the optional LDP loop detection mechanism. Its use in the Label Request message records the path of LSRs the request has traversed. Its use in the

Label Mapping message records the path of LSRs a label advertisement has traversed to setup an LSP.

The format of an exemplary CR-LDP message with the Protection TLV extension is:

```

5      0          1          2          3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
      +-+-+-+-+-+-+-+-+
|0| Label Request (0x0401)   | Message Length   |
      +-+-+-+-+-+-+-+-+
10  |      Message ID          |
      +-+-+-+-+-+-+-+-+
      |      FEC TLV           |
      +-+-+-+-+-+-+-+-+
      |      LSPID TLV         (CR-LDP, mandatory) |
15  +-+-+-+-+-+-+-+-+
      |      ER-TLV            (CR-LDP, optional) |
      +-+-+-+-+-+-+-+-+
      |      Protection TLV      (CR-LDP, optional) |
      +-+-+-+-+-+-+-+-+
20  |      Traffic TLV         (CR-LDP, optional) |
      +-+-+-+-+-+-+-+-+
      |      Pinning TLV       (CR-LDP, optional) |
      +-+-+-+-+-+-+-+-+
      |      Resource Class TLV (CR-LDP, optional) |
25  +-+-+-+-+-+-+-+-+
      |      Pre-emption TLV   (CR-LDP, optional) |
      +-+-+-+-+-+-+-+-+

```

Wherein the “Protection TLV” message field is new.

30

The Protection Object (RSVP)/Protection Type Length Value (TLV) (LDP/CR-LDP) establishes the working and a corresponding protection path utilizing the Reservation Protocol (RSVP) path message or the Constraint-Based Routing Label Distribution Protocol

RNTT= RNT Type

TO=Timer Options

RO=Recovery Option

B= Protection Bandwidth

5 RESVD=Reserved for Future Use

Since the switching systems used in the network 100 are unidirectional, and pathway fault recovery requires the notification of faults to a protection switch, such as switch no. 1 or switch
10 no. 9, responsible for a switchover to a recovery path, a mechanism is provided for the fault indication and the fault recovery notification to travel from a point of occurrence of the fault back to the protection switch. The ability to propagate a fault notice upstream however is complicated when two or more data streams merge in a single switch such as the streams from switches 9 and 2 merging at switch 3. When two or more data streams merge at a switch, e.g.
15 switch 9, a fault anywhere downstream from switch 9 will require that a fault notice be sent to multiple source switches, i.e. switches 9 and 2. The fault indication and recovery notification should be able to travel along a reverse path of the working paths to all the protection switch elements that might be affected by the fault. The path is provided by the reverse notification tree.

By way of example, Figure 2 shows an example of an MPLS protection set up
20 message sequence. The establishment of the working path and protection path is accomplished by the transmission of a Protection Switch Domain (PSD) initialization message 210 from a switch 202 to switches 204, 206 and 208. A PSD confirmation message 212 is propagated from the downstream switch 208 upstream to switch 202.

The reverse notification path, also referred to above as the Reverse Notification Tree
25 or RNT, is established by the downstream switch, 208, sending an RNT initialization message 214, upstream to switches 206, 204 and 202. Confirmation of the RNT setup is accomplished by the RNT Confirmation message 216 that originates from switch 202.

Upon the establishment of the working and protection paths, and the reverse notification tree, data 218 can be sent through the network.

30 Two "aliveness" messages 220 and 222, which provide notification of the working path status, are shown in Figure 2 to depict the fact that the aliveness message described above can be sent periodically, regardless of whether downstream data 218 was sent. As shown further, downstream data transmissions, such as transmissions 224, 226 and 228 are

not conditioned upon receipt of an aliveness message in any fixed way. An aliveness message
230 sent upstream is then followed by yet another data transmission 232.

Figure 2 shows that the sequence of aliveness messages and data transmissions do not
need to follow any sort of predetermined or fixed order. For network reliability purposes, the
5 aliveness messages are preferably sent periodically, so that their absence can be detected if
they do not arrive on time. Alternate embodiments include sending liveness messages
aperiodically.

Those skilled in the art will recognize that re-routing data on either the failure of a
link or a switch in a network such as that depicted in Figure 1 need not be performed by a
10 protection switch. In the event that switch 4 fails for example switch no. 3 might reroute data
from switch 2 that is destined for switch 7, through another protection switch element 9,
identified by reference numeral 116. Switch 9 might then reroute data from switch 2 that is
addressed to switch 7 over a protection path designed as $P_{9,10}$ and $P_{10,7}$ through switch 10,
identified as reference numeral 118.

15 In the preferred embodiment, the media over which data message are carried might be
twisted copper wires, coax cable, fiber optic cable or even a radio frequency data link. As set
forth above, each of the switching systems might accommodate a variety of packetized data
messages including but not limited to Ethernet, internet protocol, ATM, frame relay or other
types of transmission switching systems.

20 By continuously sending an upstream message indicating that downstream traffic
arrives at its destination, recovery time required to recover from the fault of a media link or a
switching system can be minimized. If the switch status message used to indicate a
functionality of a switch or a link is sent promptly enough, and to the appropriate node in a
mesh network such as that shown in Figure 1, the time required to reroute data messages
25 between first and second endpoint switches over an alternate data path can be minimized. In
the preferred embodiment, the alternate or so called protection path is preferably set up in
advance and maintained in a stand by mode such that it is immediately available when
required by the protection switch that will reroute data over the protection path.

We claim:

1. In a data network comprised of a plurality of data switches interconnected to form a plurality of data paths forming a mesh configuration of data switches, a method of re-routing data messages between first and second data switches over a pre-established alternate data path linking said first and second data switches
5 comprised of the steps of:
 - a. sending at least a first data message over a first data path from said first switch to said second switch;
 - b. receiving at said first data switch, switch status messages from said second
10 switch;
 - c. upon the loss of said switch status messages at said first switch, re-directing subsequent data messages over an alternate data path through said data network.
2. The data network of claim 1 wherein said alternate data path is a protection path
15 through said network.
3. The method of claim 1 wherein said data switches are asynchronous transfer mode switches.
4. The method of claim 1 wherein said data switches are internet protocol (IP) routers.
- 20 5. The method of claim 1 wherein said switch status messages are comprised of a predetermined format, switch liveness message.
6. The method of claim 1 wherein at least one of said switches maintains a table of incoming link and path identifiers and of outgoing link and path identifiers.
7. The method of claim 1 wherein said first data message represents speech
25 information.
8. The method of claim 1 wherein said first data messages represents computer data.

9. The method of claim 1 wherein said step of re-directing said series of data messages from said first path over another path through said data network includes the steps of:

sending subsequent first data messages to a third data switch.

10. The method of claim 1 wherein said first data switch is a protection switch element.
11. In a data network comprised of a plurality of data switches interconnected to form a plurality of data paths forming a mesh configuration of data switches, a method of re-routing data messages around a data switch comprised of the steps of:
 - a. sending at least a first data message over a first data path from a first switch to a second switch;
 - b. sending said at least a first data message from said second switch to a third switch;
 - c. receiving at said second data switch, switch status messages indicating the functionality of said third data switch;
 - d. upon the loss of said switch status messages at said second switch, sending a switch failure message from said second switch to said first switch;
 - e. upon the receipt of said switch failure message at said first switch, said first switch re-directing subsequent data messages away from said second and third switch via a second data path through said data network.
12. The data network of claim 11 wherein said second data path is a protection path through said network.

13. The method of claim 11 wherein said data switches are asynchronous transfer mode switches.
14. The method of claim 11 wherein said data switches are internet protocol (IP) routers.
- 5 15. The method of claim 11 wherein said data switches are digital cross connect switches controlled by MPLS.
16. The method of claim 11 wherein said data switches are optical cross connects and switches controlled by MPLS.
- 10 17. The method of claim 11 wherein said switch status messages are comprised of a predetermined format, switch liveness message.
18. The method of claim 11 wherein at least one of said switches maintains a table of incoming link and path identifiers and of outgoing link and path identifiers.
19. The method of claim 11 wherein said series of data messages represent speech information.
- 15 20. The method of claim 11 wherein said series of data messages represent computer data.
21. The method of claim 11 wherein said step of re-directing said series of data messages from said first path over another path through said data network includes the steps of:
20 sending subsequent data messages to a third data switch.
22. In a data network comprised of a plurality of data switches interconnected to form a plurality of data paths forming a mesh configuration of data switches, a method of re-routing data messages between first and second data switches over a pre-established alternate data path linking said first and second data switches
25 comprised of the steps of:

- a. sending at least a first data message over a first data path from said first switch to said second switch;
- b. upon the loss of said first data message at said second switch, sending a switch status messages to said first switch, the receipt of said switch status message thereby causing the re-directing of subsequent data messages over an alternate data path through said data network.

23. In a data network comprised of a plurality of data switches interconnected to form a plurality of data paths forming a mesh configuration of data switches, a method of re-routing data messages around a data switch comprised of the steps of:

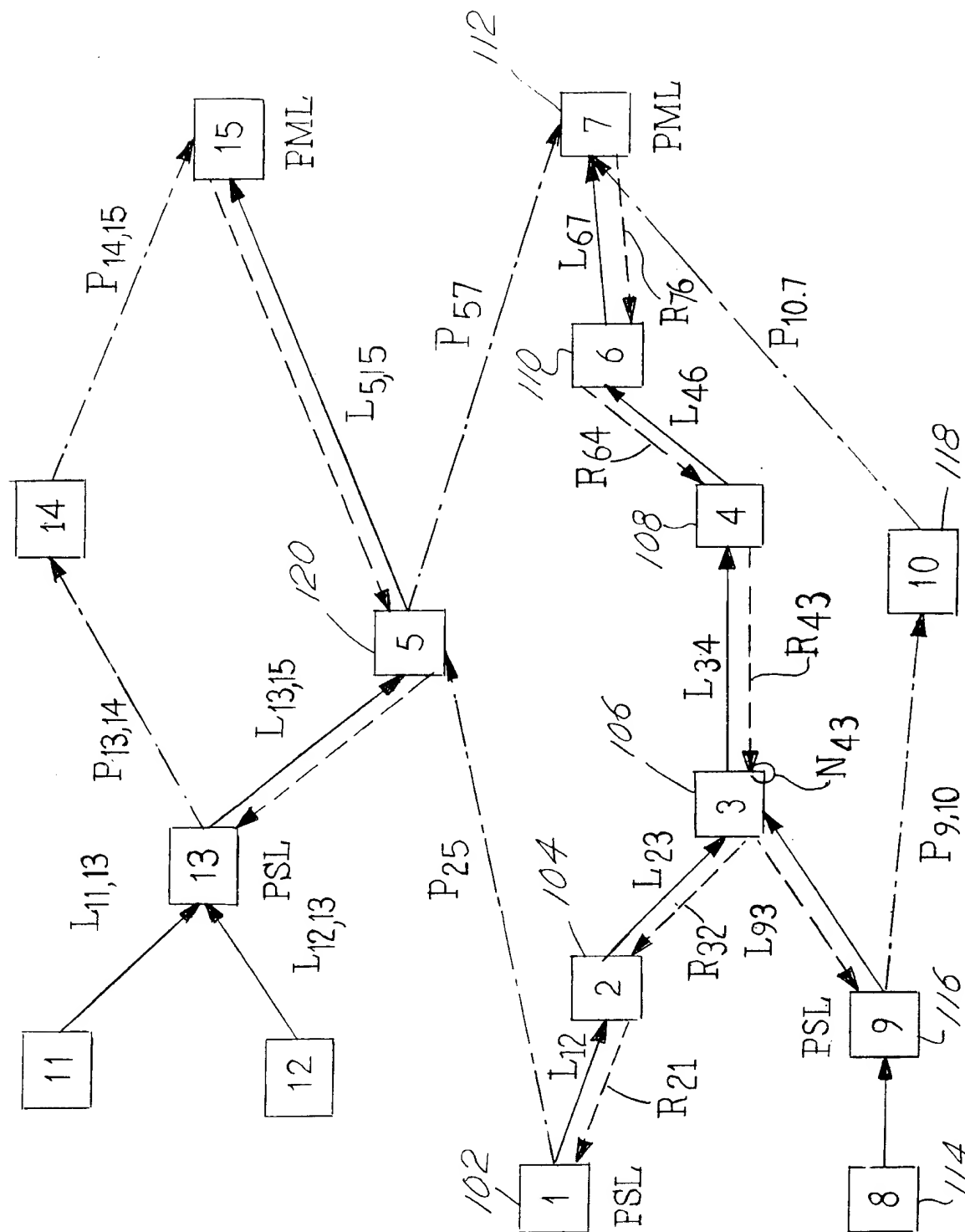
5 a. sending at least a first data message over a first data path from a first switch to a second switch;

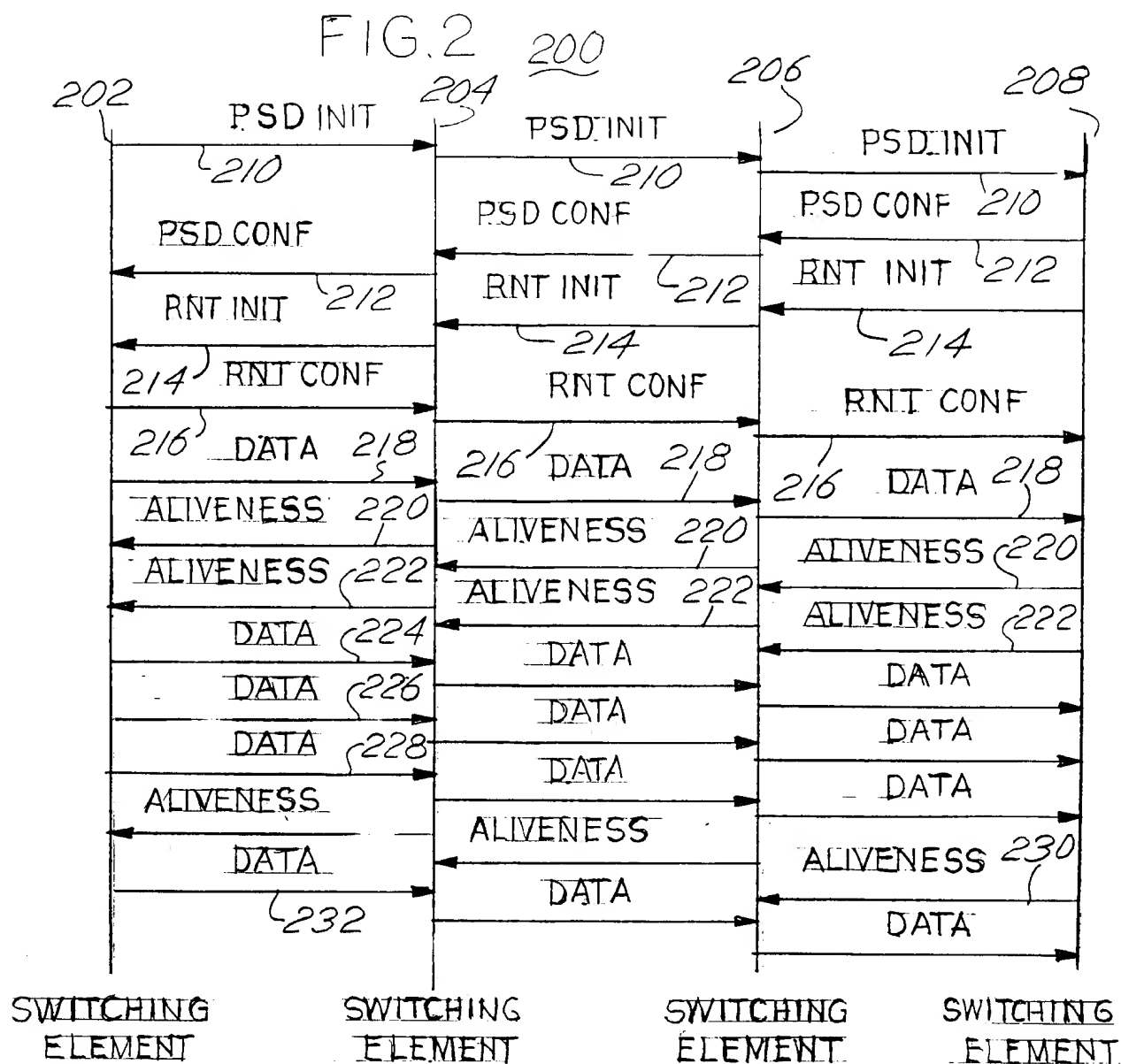
b. sending said at least a first data message from said second switch to a third switch;

10 upon the loss of said first data message at either said second switch or said third switch, sending a switch status message to at least one of said first and second switches thereby causing the re-directing of subsequent data messages away from said second and third switch via another data path through said data network.

24. The method of claim 23 wherein said first switch is a protection switch element.

FIG. 1





PSD=PROTECTION SWITCH DOMAIN(ESTABLISHMENT OF THE WORKING AND PROTECTION PATH: IDENTIFICATION OF SWITCHING ELEMENTS)

RNT=REVERSE NOTIFICATION TREE (ESTABLISHMENT OF REVERSE PATH FOR NOTIFICATION OF ALIVENESS)

INIT=INITIALIZE

CONF=CONFIRM

DATA=DATA FLOW

ALIVENESS=NOTIFICATION OF WORKING PATH STATUS

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/28992

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 15/173

US CL : 709/238, 239

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 709/238, 239; 714/2, 4; 370/217, 244, 245

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

STN

search terms: alternative path, ATM, MPLS, acknowledgement, fault recovery

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,450,578 A (MACKENTHUM) 12 September 1995, see entire document.	1-24
X	US 5,771,227 A (BENAYOUN et al.) 23 June 1998, col. 3, line 52 - col. 9, line 19.	1-24
A	US 5,838,924 A (ANDERSON et al.) 17 November 1998, col. 4, line 15 - col. 7, line 20.	1, 11, 22, 23
A	US 5,991,300 A (TAPPAN) 23 November 1999, see entire document.	1, 11, 22, 23
A	US 6,055,561 A [FELDMAN et al.] 25 April 2000, col. 7, line 8 - col. 13, line 44.	1, 11, 22, 23
A,P	US 6,065,061 A [BLAHUT et al.] 16 May 2000, see entire document.	1, 11, 22, 23

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A" document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance, the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E" earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O" document referring to an oral disclosure, use, exhibition or other means		
"P" document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

04 JANUARY 2001

Date of mailing of the international search report

26 FEB 2001

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

MARK RINEHART

Telephone No. (703) 305-3800

Peggy Harrod